图书情报工作
第 62 卷 第 13 期 2018 年 7 月

一种层次化的科学知识结构发现方法*

■ 李慧 田亚丹
西安电子科技大学经济与管理学院 西安 710126

摘要：[**目的/意义**] 提出一种新的层次化科学知识结构发现方法，为优化知识结构发现过程，改善知识组织形式提供借鉴。[**方法/过程**] 利用 LDA 主题模型构建层次化的科学知识结构发现方法，依据主题间平均相似性自动确定知识结构层数，通过在"文档－主题"概率矩阵中自动筛选阈值截取各主题文献子集，最后采用树形图展示科学领域的知识结构，发掘知识间的关联性和继承性，并与层次主题模型 HLDA 方法进行比较。[**结果/结论**] 通过实证研究与对比，证明本文提出的方法得到的知识结构更优，知识主题表征性更强且运行效率更高，并在单层主题区分度和层间主题继承性方面较 HLDA 方法有较大提升。

关键词：LDA 云计算 层次化 知识结构
分类号：G254
DOI：10.13266/j.issn.0252－3116.2018.13.012

1 引言

科学研究的泛化、交叉和渗透使各领域研究呈现出交错复杂的局面，研究内容的多样化为理解与掌握知识内在结构带来一定困扰，浩瀚的知识和有限的个人精力之间的矛盾不可避免。对新接触某领域的学者来说，想要全面了解该领域知识结构往往需要很多工作量[1]。分散的知识点、非结构化的信息不利于知识结构的形成，也阻碍了学科的深度交叉与融合。因此，构建一个科学合理的知识结构对科学研究具有重要意义。而目前有关知识结构的研究，多以文献计量学为基础，采用多元统计分析或社会网络分析的方法，利用关键词共现等对知识点进行简单抽取，侧重科学热点的发现，无法完整揭示科学研究的内在知识结构。

针对传统研究方法的不足，本文提出一种新的层次化科学知识结构发现方法，该方法利用 LDA 主题模型生成的特征词对知识点进行潜在映射，挖掘其更深层的语义信息，突破了传统方法仅根据共现词刻画主题热点的局限性，揭示了科学领域知识点间的多粒度层次关系，解决了知识结构发现中常常忽略知识间继承性的问题，更能体现知识组织的本质。研究者可以通过浏览层次化的知识结构全面、快速地了解领域概况和知识点分布情况，有利于减少阅读的工作量。

2 相关研究

早期关于知识结构的研究多是从"人才培养"的角度出发，讨论教师、图书馆员或各领域工作人员应该具备怎样的知识结构[2]。2000 年后，有关知识结构的研究开始将关注点转向科学文献，涉及领域也愈加广泛。目前，国内外有关知识结构发现的研究方法大致分为以下三类：基于多元统计分析的方法；基于社会网络分析的方法和基于主题模型的方法。

基于多元统计分析（Multivariate Statistical Analysis，MSA）的知识结构发现方法主要对研究领域进行分类或预测其发展趋势，已被成功地应用于供应链[3]、知识管理[4]等领域。用于知识结构发现的 MSA 方法从经典统计学理论发展而来，常通过多维标度法（Multidimensional Scaling，MDS）绘制二维知识地图，再借助聚类分析（Cluster Analysis，CA）或因子分析（Factor Analysis，FA）的结果确定知识点的数目和边界[5]。其

* 本文系国家自然科学青年基金项目"基于可信语义 Wiki 的知识库构建方法与应用研究基金"（项目编号：71203173）、国家自然科学青年基金项目"大规模动态社交网络社团检测算法研究"（项目编号：71401130）和中央高校基本科研业务费专项资金资助项目"大数据环境下基于主题模型的信息服务研究"（项目编号：JB160606）研究成果之一。
作者简介：李慧（ORCID：0000－0002－3468－5170），副教授，博士，硕士生导师；田亚丹（ORCID：0000－0001－8506－8271），硕士研究生，通讯作者，E-mail：179075482@qq.com。
收稿日期：2018－01－17 修回日期：2018－04－01 本文起止页码：92－102 本文责任编辑：杜杏叶

李慧，田亚丹. 一种层次化的科学知识结构发现方法[J]. 图书情报工作,2018,62(13):92 - 102.

中,MDS 可以通过非线性变换将知识点映射到低维空间,CA 可以用谱系图的方式展示知识间的关系,而 FA 主要是用较少的因子替代所有变量从而简化分析过程。大多数研究都是将三种方法进行综合利用来发掘领域知识结构。

基于社会网络分析(Social Network Analysis,SNA)的知识结构发现方法起初以文献计量学为基础,利用文献外部特征,如作者、引文、机构等分析科学领域的合作关系[6-7]和引用关系[8],研究成果十分丰富。随着研究的深入,从文献内部特征如摘要、关键词等入手,利用 SNA 方法发现科学文献知识结构的研究也逐步展开,已应用在西方经济地理学[9]、图书馆学与情报学[6]等领域。国外学者还将 MSA 和 SNA 的方法结合起来,互为补充,进行知识结构发现工作[10-11]。

上述两类方法的共同点在于通过构建关键词共现矩阵进行聚类,以二维地图或网络图的形式对知识点进行可视化展示,发掘核心研究群体,追踪领域发展脉络,侧重于学科前沿、热点的研究。此外,MSA 方法对知识的整体属性和节点间的联系不敏感,容易忽视特殊节点和小知识群。而 SNA 方法易受软件固有功能的影响,数据转换时可能丢失有用的信息,数据规模有一定限制,且整个过程中人为干预较多。这两种方法得到的知识结构均难以反映知识间更深层次的语义关系,而主题模型的出现可以很好的解决上述问题。

基于主题模型的知识结构发现方法主要采用 LDA (Latent Dirichlet Allocation)模型[12]或 LDA 的改进模型。此类方法一般利用科学文献标题、摘要等构建语料库,通过主题模型抽取潜在的知识主题及表征其内容的特征词。2016 年王曰芬等从学科分类的角度出发,采用 LDA 主题模型深入挖掘国内知识流领域的知识结构[13],同年, H. C. Chang 利用 LDA 主题模型挖掘信息安全领域的知识结构[14]。由于通过 LDA 模型得到的知识主题是由一系列特征词组成的,这种描述方式可以反映主题的语义信息,每篇文档属于各主题的分布情况也可以通过概率直观的展示出来,有效避免了 MSA 和 SNA 方法仅通过共现词对刻画知识主题的单一性,也更适用于大规模知识主题的抽取工作,但上述方法大多直接利用 LDA 经典模型进行主题抽取,其结果依然是表层的知识热点,没有对知识结构层次化的完整表示,忽略了知识间的层次继承关系。针对上述问题,有学者开始利用 LDA 的拓展模型层次主题模型 HLDA(Hierarchical Latent Dirichlet Allocation)[15]进行知识组织的研究工作,在图书内部主题组织[16]和专利分析[17]上都有应用。不过此方法得到的层次主题结构依然存在一些问题,如较多的主题重叠现象、文档在叶子节点中的分布较为稀疏、层次结构和主题区分度不好控制等。

因此,为了更好的进行层次化科学知识结构发现工作,改变 MSA、SNA 等传统知识结构发现方法仅挖掘科学领域的表层知识热点,难以反映知识间层次继承关系的情况,针对 HLDA 方法的不足之处,本文重点关注层次化科学知识结构发现中的两大问题,一是如何确定知识结构的层数,以保证层次粒度的合理划分,二是如何确定文献子集的范围,使抽取的知识主题更准确。本文在设计模型时,一方面,利用知识主题间的相似性,为确定知识结构层数提供合理依据,另一方面,从生成知识主题的初始文献集入手,平衡文档范围和文献质量的关系,最大程度提升主题质量,优化知识结构发现过程,并与 HLDA 这一经典的层次主题模型进行对比,比较各自的优劣,为知识结构发现的相关研究方法提供参考。

3 层次化知识结构构建

3.1 研究框架

本文基于 LDA 主题模型进行层次化知识结构发现,其研究框架如图 1 所示,自上而下分为数据层、逻辑层和展示层三部分:数据层是文献收集与预处理过程,主要对文献语料进行初步处理,包括对收集到的文献集进行合并整理,以及分词、去除停用词等文献预处理的一般流程;逻辑层是对层次化科学知识结构的发现过程,利用 LDA 模型对预处理好的语料进行挖掘,并通过计算主题间平均相似性帮助确定知识结构层数,通过自动筛选阈值帮助确定下层主题的文献子集范围;展示层主要根据逻辑层得到的特征词对知识主题进行映射,绘制科学知识结构树形图,便于学者快速了解领域知识及其分布情况。

3.2 文献收集与预处理

文献收集首先要确定检索式,然后在专业的数据库中检索文献并导出所需记录,最后对文献数据进行合并汇总,删除题录信息(包括标题、摘要等)不完整的文献。对于主题抽取工作,原始数据的质量是很重要的,收集到的文献数据是非结构化的,在抽取主题之前要统一进行预处理,包括分词,去除停用词和词形还原,从而生成构建层次化科学知识结构所需的建模文件。建模文件的第一行是语料库中的文档总数,每个文档占据一行。由于本文采用中英两种文献集进行分

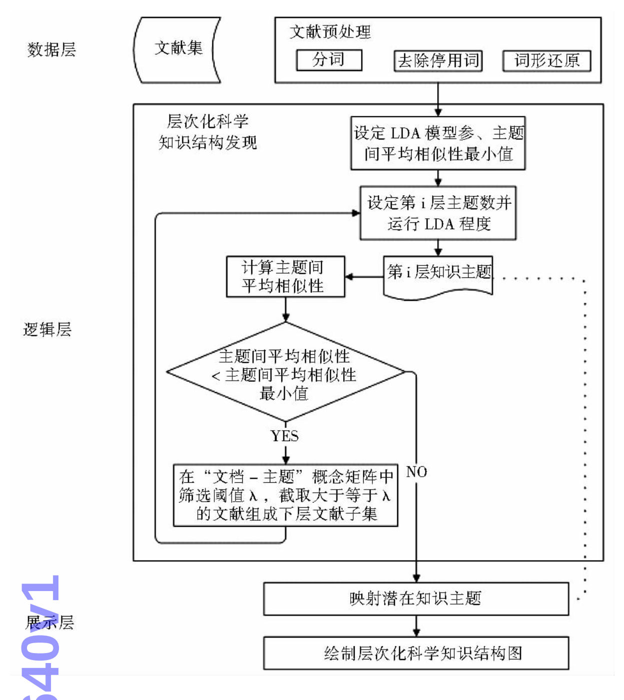


图 1　层次化科学知识结构发现框架

析，在预处理操作时略有不同，如英文不需要分词，中文不需要词形还原。

3.3　层次化科学知识结构发现方法

逻辑层是层次化科学知识结构发现方法的核心部分，具体实现流程如下：

Step1：将预处理好的建模文件输入 LDA 模型，设定 LDA 模型参数（参见 4.2.1 节）和主题间平均相似性最小值；

Step2：设定第 i 层主题数，运行 LDA 程序输出“文档－主题”概率分布和“主题－词”概率分布等文件，由此得到第 i 层知识主题；

Step3：计算主题间平均相似性，判断该值与之前设定的主题间平均相似性最小值的大小，从而确定知识结构层数，若判断结果为小于，执行 Step4，否则转向 Step6，其算法描述见 3.3.1 节；

Step4：继续层次化过程，为下层主题重新截取文献子集，其算法描述见 3.3.2 节；

Step5：令 i＝i＋1，转向 Step2，生成下层知识主题；

Step6：此时已满足层次终止条件，为了保证主题间良好的区分性，不再挖掘下层主题，至此对全部潜在知识主题进行映射，绘制层次化的科学知识结构图。

3.3.1　确定知识结构层次　随着知识结构层次化的加深，知识粒度逐渐细化，同一主题下的知识点越来越相似，此时继续进行下一层主题的抽取便没有意义，所以本文设计了一个确定知识结构层次的算法，这里，需事先设定一个主题间平均相似性最小值 S_0（根据经验设定在 0 到 1 之间即可）。

该算法基于比较成熟的向量空间模型，其实质是将主题下的特征词映射到向量空间，获得各主题的特征向量，具体实现方法如下：对于第 i 层主题 T_{ik} 和 T_{ir}，$(1 \leqslant k \leqslant j, 1 \leqslant r \leqslant j, k \neq r)$ 其中 k 和 r 为主题编号，j 代表主题总量，分别将其下前 25 个特征词不重复的合并为一个集合，（将主题下特征词按分布概率降序排列，前 25 个特征词可以较好的表征主题，故本文各主题下特征词均设为 25 个。）可以得到一个 n 维向量，n 是特征词总数，每个主题由一个向量 $V[v_1, v_2, \cdots, v_n]$ 表示，V_a 是该主题对于第 a 个特征词的分布频次。如果该主题不包含特征词 a，则 $V_a = 0$，至此，我们得到主题的特征向量 V。

计算两个主题特征向量的夹角余弦得到主题间相似性[18]，最后遍历该层所有主题，计算主题间相似性的均值并与设定值 S_0 进行比较，从而确定终止层。整个算法的伪代码如图 2 所示：

```
输入：第 i 层所有主题下的特征词 Ti_word
输出：终止层数
class HierarchyDetermine
 method get (Vik,Vir)  //生成主题 Tik 和 Tir 的特征向量
   for all topic [Tik,Tir] do  //Tik ∈ [Ti1, Ti2, …, Tij], Tir ∈ [Ti1, Ti2, …, Ti(k-1), Ti(k+1), …Tij]
     V = [v1, v2, … va, …, vn]  //将主题下的特征词映射到向量空间
   end for
 method similaritycomputation (Vik,Vir)
   for all V [Vik,Vir] do
     S_i,kr = (Vik • Vir) / (||Vik|| × ||Vir||)  //计算主题间相似性S_i,kr
   end for
 method sum (S_i,kr)
   for all [ S_i,k1, S_i,k2, …, S_i,k(k-1), S_i,k(k+1), …, S_i,kj ] do
     S_ik = Σ_{r=1}^{k-1} S_i,kr + Σ_{r=k+1}^{j} S_i,kr  //计算主题 Tk 与该层其他各主题的相似性之和S_ik
   end for
 method avgsimilaritycomputation(S_ik)
   for all [S_i1, S_i2, …, S_ik, …, S_ij] do
     // S_i1, S_i2, …, S_ik, …, S_ij为每个主题与该层其他各主题的相似性之和
     Avg(S_i) = 2 Σ_{k=1}^{j} S_ik / (j(j-1))  //计算第 i 层的主题间平均相似性 Avg(S_i)
   end for
 method judge( Avg(S_i), S_0)  //判断 Avg(S_i)和 S_0 的大小
   for all hierarchy [1,……,i] do
     if Avg(S_i) < S_0
       do ThresholdDetermine  //执行 3.3.2 中的算法，确定下层文献子集范围
       i=i+1  //继续层次化过程，运行 LDA 程序生成下层知识主题
       get (Vik,Vir)
       similaritycomputation (Vik,Vir)
       sum (S_i,kr)
       avgsimilaritycomputation(S_ik)
       judge( Avg(S_i), S_0 )
     else besthierarchy=i  //i 即为知识结构终止层
     end if
   end for
```

图 2　确定知识结构层次伪代码

3.3.2　确定文献子集范围　在主题模型中，每篇文档都以一定的概率归属于各个主题，要对主题进行细化，刻画其下层知识结构，就要对属于该主题的文献再次

进行主题抽取，此时文献子集的质量会对下层主题的抽取效果产生很大影响。传统方法一般根据经验设定一个数值或百分比来截取文献集，此方法带有很大的主观性，若选择排序靠前的大概率文档，下层主题对上层主题的继承性较好，但可能会丢失小概率文档产生的重要主题，若将文档范围扩大，则会造成文献集质量下降，程序运行效率降低，同时下上层知识主题很相似，不能达到细化的目的。

因此，为了在文档范围和主题质量这两方面找到一个平衡点，通过观察"文档－主题"概率矩阵，我们发现，如果可以找到一个阈值 λ，使之满足以下两个条件：①在重复分配率较小的情况下，保证所有文档都可以被分配到相应的主题下，不会因为截取阈值太小而造成文档丢失，②被分配到各主题下的文档都对该主题有较高的归属概率，不会因为截取阈值太大而影响主题质量。那么这样的阈值就是我们要找的平衡点。

以图 3 为例，在一个 6×5 的"文档－主题"概率分布矩阵中，我们首先筛选出每一行的最大值，分别是 0.275、0.3889、0.9198、0.6868、0.1328 和 0.4841，再筛选这六个值当中的最小值 0.1328，以该值作为截取文献集的阈值 λ（红色标记），使被截取的文献在相应主题下的分布概率较高，使主题质量得到保证，满足条件②。自动筛选出阈值 λ 后，我们为每个主题截取概率大于等于 λ 的文档作为该主题的下层文献子集。截取结果（绿色标记和红色标记）为 Topic0 下有文档 3 和 4；Topic1 下有文档 1、2 和 6；Topic2 下有文档 4；Topic3 下有文档 5；Topic4 下有文档 3。这样在不丢失任何文档的前提下，6 篇文档以较低的重复分配率归属于各主题下，满足条件①。

主题 文档	Topic0	Topic1	Topic2	Topic3	Topic4
文档 1	0.025	0.275	0.025	0.0083	0.025
文档 2	0.0079	0.3889	0.0238	0.0238	0.0238
文档 3	0.9198	0.0062	0.0062	0.0062	0.2062
文档 4	0.6868	0.0055	0.2374	0.0385	0.0495
文档 5	0.0078	0.0078	0.0078	0.1328	0.0078
文档 6	0.0079	0.4841	0.0079	0.0397	0.0079

图 3　文档－主题概率分布矩阵（部分）

综上，对于一个"文档－主题"概率分布矩阵，我们首先筛选出每一行的最大值，再筛选出这些值中的最小值，该值即可作为下层主题文献子集的截取阈值。用 $P_{(D_{im},T_{ik})}$ 表示第 i 层主题中文档 m 隶属主题 k 的概率，其中 m 为文献编号，1≤m≤w，w 为文献总量，像这样通过两次取最值得到阈值 λ 的方法可用符号表示为：$\lambda = \wedge(\vee_w P_{(D_{im},T_{ik})})$（符号 ∧ 代表"取小"运算，∨ 代表"取大"运算）。其算法的伪代码如图 4 所示：

```
输入："文档—主题"概率矩阵 P
输出：文献子集
class ThresholdDetermine
  method rowmax   //在概率矩阵 P 中自动筛选每一行的最大值
    for all row [1,2,…, m,…, w] do
      λ_w = (∨_w P_(D_im,T_ik))   //得到每一行的最大值
    end for
  method colmin   //在得到的 w 个最大值中自动筛选一个最小值
    for all [λ_1, λ_2,……, λ_m,……, λ_w] do
      λ = ∧(∨_w P_(D_im,T_ik))   //得到阈值λ
    end for
  method judge   //判断文档概率与阈值λ的大小
    for all topic (T_i1, T_i2,…, T_ik,…, T_ij) do
      sort P_(D_im,T_ik)
      while P_(D_im,T_ik) ≥ λ do
        addInNextArticleSubsets
        //为各主题截取概率大于等于阈值λ的所有文献形成该主题的下层文献子集
      end while
    end for
```

图 4　确定文献子集范围伪代码

3.4　生成知识结构层次树

展示层主要通过图形化的界面向用户呈现最终的层次化科学知识结构，利用主题下的特征词映射潜在知识主题，通过对该领域较为了解的专家总结出符合实际的主题代表词，帮助学者更好地理解主题语义信息，根据总结好的主题词逐层绘制科学知识结构，对科学知识结构进行层次化的完整表示，生成知识结构层次树。这种层次结构可以清楚的显示知识点间的并列关系和继承关系，粗粒度与细粒度的划分，更有利于学者对科学领域知识的全面了解。

4　实验

"云计算"（Cloud Computing）一词自提出以来受到了学界、业界的广泛关注，也是近几年国内外的研究热点之一。为了全面探究该领域发展状况，帮助研究者了解该领域的知识结构，同时对本文提出的层次化科学知识结构发现方法（用 HSKSD 表示，Hierarchical Scientific Knowledge Structure Discovering Method）进行验证，我们选取云计算领域的中英文文献作为实验数据来源，挖掘知识主题，绘制层次化的知识结构，最后分析实验结果并与经典的 HLDA 方法进行比较。

4.1　数据概况与预处理

4.1.1　数据概况　本文按表 1 所示进行检索，过滤掉少量题录信息不完整的文献，最终得到中文文献记录 6 115 条，英文文献记录 4 843 条。

表 1 数据获取

数据类型	检索时间范围	数据库来源	检索表达式	来源类别	文献类型	检索结果
中文文献	2005.01.01－2016.12.1	中国知网 CNKI	SU＝“云计算”	SCI 来源期刊、EI 来源期刊、核心期刊和 CSSCI	期刊	6 115 篇
英文文献		Web of Science 核心合集引文数据库	TS＝“cloud computing”	SCI-EXPANDED、SSCI、CPCI-S 和 CPCI-SSH	Article	4 843 篇

如图 5 所示，2005 年和 2006 年的中英文均没有检索到任何文献，说明“云计算”的发展正式起源于 2007 年。截至 2016 年，该领域历经近十年的发展，其发文趋势保持稳步上升，中文在 2013 年，英文在 2015 年的发文量均突破 1 000 篇，可见云计算领域的影响力正在逐步扩大，且外文研究在 2016 年后有赶超中文研究的趋势。

4.1.2 数据预处理 预处理过程是对原始文献语料的加工，生成建模所需的数据文件。针对不同数据类型的文献集，我们采用不同的预处理方法，具体操作见表 2。

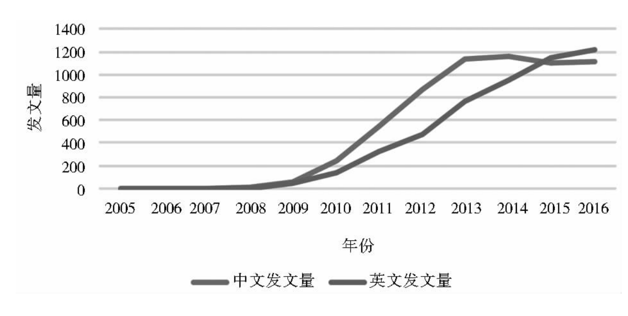


图 5 云计算领域历年发文趋势图

表 2 数据预处理

数据类型	预处理方法	词性还原	去除停用词	处理内容
中文文献	Hanlp 软件包[19]	不需要	百度和哈工大的停用词表，加之自定义词汇	标题，摘要
英文文献	斯坦福 coreNLP 软件包[20]	需要	Lucene 停用词表	标题，摘要

4.2 实验设置与结果分析

4.2.1 实验设置 由于 LDA 的相关算法已经比较成熟，在此不做过多说明，其参数设定参考文献[21]和实验总结的经验值，α 和 β 用来控制主题和词语的分布，具体说明如表 3 所示：

表 3 LDA 建模参数说明

模型参数	参数说明
α	文本集在潜在主题上的狄利克雷先验，α＝0.5
β	潜在主题在特征词集上的狄利克雷先验，β＝0.02
T_i	各层潜在主题数，$T_1=1$，$T_2=10$，$T_3=6$
niters	Gibbs 抽样迭代次数，niters＝1000
twords	主题下特征词个数，twords＝25

HLDA 建模步骤及参数设定参考文献[22]，同时为了方便后续评估，使两种方法生成的知识结构层次和主题数量尽量接近，经过多次试验，具体参数设定如表 4 所示：

表 4 HLDA 建模参数说明

模型参数	参数说明
α	文本集在潜在主题上的狄利克雷先验，α＝20
gamma	nCRP 参数[23]，决定先验树结构的形状，即文档每一层的路径选择，gamma＝20
eta	狄利克雷分布超参，即每一层的主题参数，eta＝0.09

对于云计算的中文、英文文献数据集，完成相关题录信息的预处理后，将建模文件输入基于 Java 的 LDA 程序，参照 3.3 节层次化科学知识结构发现方法进行实验。为得到结构较优的层次化知识结构，经多次试验，我们选取表 5 中的阈值 λ 和主题间平均相似性最小值 S_0 作为比较参数。

表 5 比较参数设定值

比较参数	截取阈值 λ	主题间平均相似性最小值 S_0
中文	0.19	0.25
英文	0.17	0.27

4.2.2 实验结果分析 根据主题下特征词的分布情况，逐层映射潜在知识主题，得到云计算领域的三层知识结构，其中第二层知识主题及其词语概率分布情况的部分结果展示如下表 6、7 所示。将特征词按概率分布降序排列，可以看到，同一主题下的特征词有较大相关性，它们所表达的语义信息也较为接近，这对于潜在知识主题的映射有很大帮助。我们分别绘制 HSKSD 和 HLDA 方法得到的“云计算”领域中英文语料下的知识结构层次树，展示如图 6－9 所示。

对 HSKSD 方法生成的知识结构进行解读，中英文的首层知识主题均为“云计算”，中文第二层的十大主题包括算法优化、教育领域、用户服务、数据安全、产业创新、存储处理、技术研究、智能检测、信息服务和平台架构，英文第二层的十大主题包括 Mobile network、Algorithm scheduling、Virtual machine、Image system、Cloud

表 6　中文第二层知识主题及其词语概率分布情况(部分)

主题 1 算法优化		主题 2 教育领域		主题 3 用户服务		主题 4 数据安全	
资源	0. 04453305	教育	0. 03449461	服务	0. 08079356	安全	0. 04778969
算法	0. 04347546	学习	0. 03192548	模型	0. 03641950	方案	0. 02981086
调度	0. 03182569	云计算	0. 02388114	云计算	0. 02492295	用户	0. 02814273
云计算	0. 02655400	技术	0. 01638433	用户	0. 02410365	数据	0. 02180998
虚拟机	0. 01861393	教学	0. 01566834	方法	0. 01805144	云计算	0. 02165552
优化	0. 01853258	中国	0. 00977197	环境	0. 01662428	环境	0. 02125393
负载	0. 01671027	实践	0. 00926657	研究	0. 01535570	存储	0. 01856637
策略	0. 01558759	专业	0. 00762401	提出	0. 01511784	加密	0. 01624951
提出	0. 01526218	环境	0. 00728707	信任	0. 01334710	保护	0. 01597149
……		……		……		……	

表 7　英文第二层知识主题及其词语概率分布情况(部分)

主题 1 Mobile network	主题 2 Algorithm scheduling	主题 3 Virtual machine	主题 4 Image system
network0. 08469845	algorithm0. 03492635	virtual0. 04148223	system0. 03544545
mobile 0. 05664680	cloud0. 03223510	performance0. 03571050	image0. 01819276
cloud 0. 02924265	energy0. 02924316	machine0. 03258906	health0. 01805041
computing0. 02512785	scheduling0. 02333444	cloud0. 02874124	cloud 0. 01651304
device 0. 02471010	resource0. 02183096	system 0. 065297	user 0. 01449168
propose0. 01330563	computing0. 02109425	virtualization0. 01815974	video 0. 01446321
internet 0. 01297143	task 0. 01871873	server 0. 01798305	content 0. 01229951
communication 0. 01286700	time 0. 01715511	application 0. 01778673	multimedia 0. 01039204
datum 0. 01269990	optimization 0. 01641840	resource 0. 01733520	social 0. 01033510
……	……	……	……

application service、Datum security、Computing method、User resource service、Business management 和 Data analysis。查阅资料，我们发现，这些知识主题与云计算领域的专业图书[24]中介绍的相关概念与技术不谋而合，与前人在"云计算"领域的相关研究结果也基本一致[25-26]，证明了本文的主题抽取结果具有一定的科学性和合理性。分析知识结构中的第三层，可以发现，除了对传统技术，如算法、存储、数据安全和平台架构等的持续关注外，近些年来，中英文云计算领域的研究还涉及到教育、金融、交通和政务等各个方面，其应用范围愈加广泛。我们知道，云计算实际上是一组相关技术和服务的总称，与云计算领域的技术和服务相结合，各种新兴技术如物联网技术、图像检索技术等受到学者的广泛关注，各类特色服务如推荐服务、智能监测服务等不断涌现。值得一提的是，图情领域也因此发生了一些变革，如图书馆、出版业和媒体业的转型，涉及知识共享、知识管理以及文献分析的相关研究越来越多。

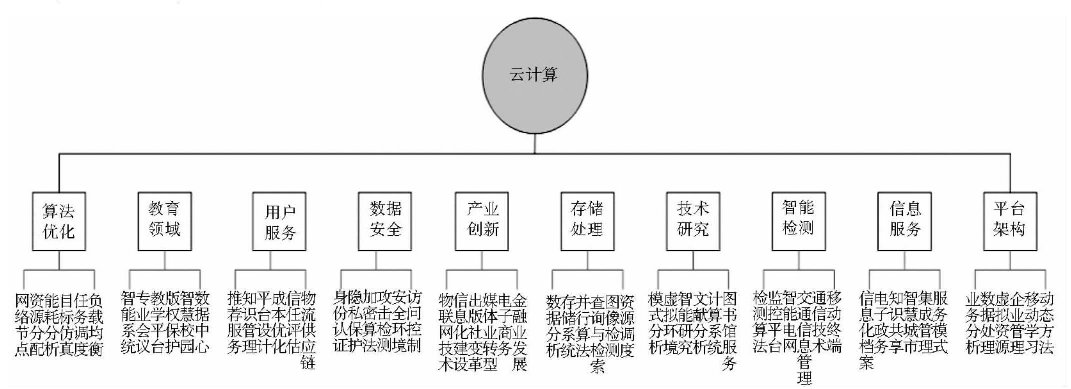


图 6　HSKSD 方法——中文层次化知识结构

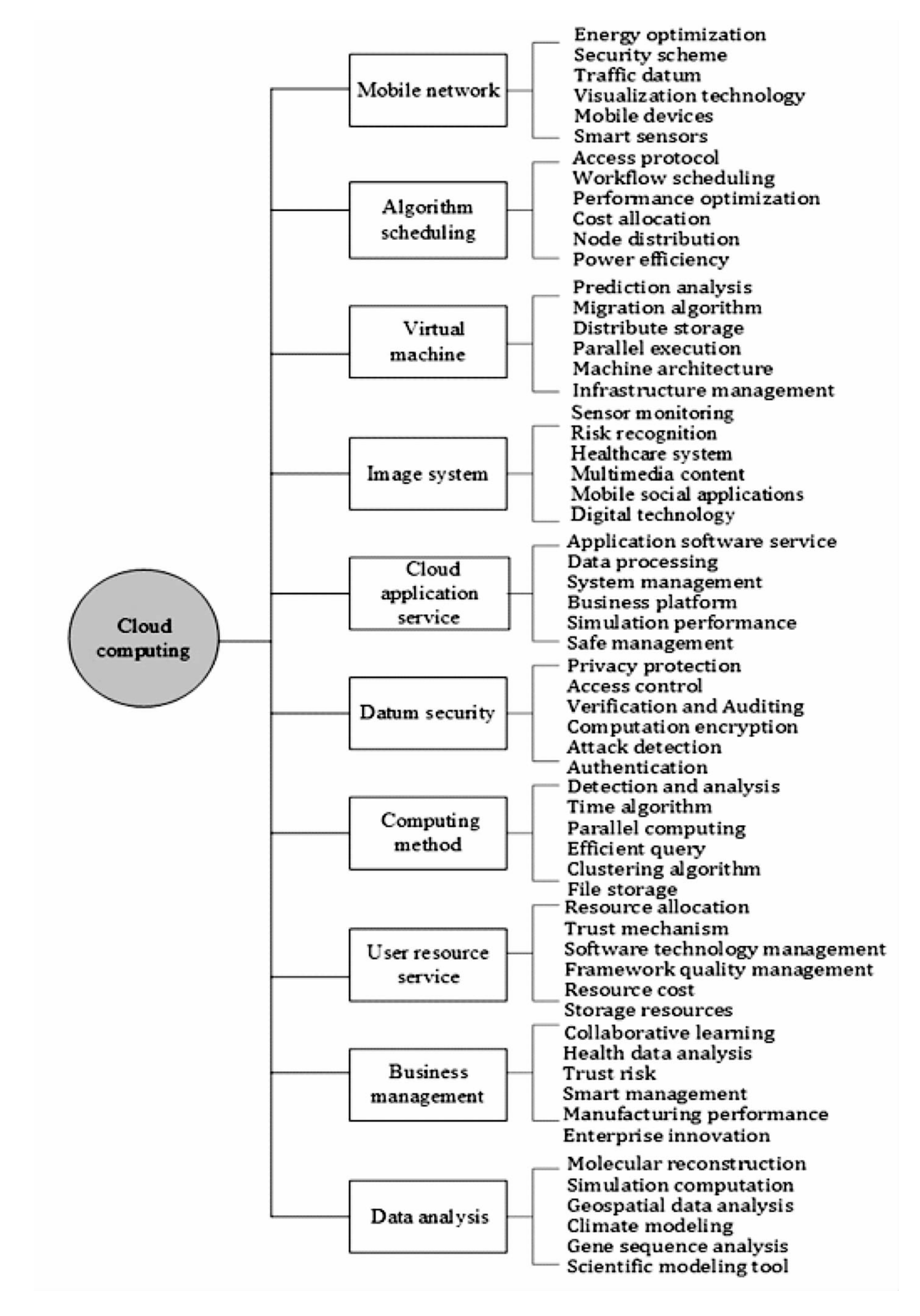


图 7 HSKSD 方法——英文层次化知识结构

对于 HLDA 方法生成的知识结构进行解读,中英文的首层知识主题依然为“云计算”,中文第二层知识主题包括物联网、虚拟机、服务器、数据库、资源共享、基础设施、教育领域、移动通信、空间信息、数据源、服务外包共十一个主题。英文第二层知识主题包括 system datum、mobile datum、computing performance、distributed database、network system、distribution strategy、detection analysis、time device 共八个主题。涉及云计算在教育、通信等领域的应用。另外,分析其生成的第三层知识主题,“城市规划”和“遗传信息”展示了云计算技术在智慧城市和生物医药领域的贡献,但总体而言,其主题间的从属关系较为混乱,如“虚拟机”下的“知识产权”,“移动通信”下的“遗传信息”,英文也存在同样的问题,如“computing performance”下的“business innovation”和“privacy protection”。

对上述结果从以下两方面进行定性分析:

(1)单层主题区分度。以中文第二层知识主题为例,对比两种方法生成的知识主题可以发现,HSKSD 方法得到的知识主题分类更加清晰,知识间区分度强,对应的特征词有较准确的表征,基本涵盖了云计算领

李慧，田亚丹. 一种层次化的科学知识结构发现方法[J]. 图书情报工作,2018,62(13):92－102.

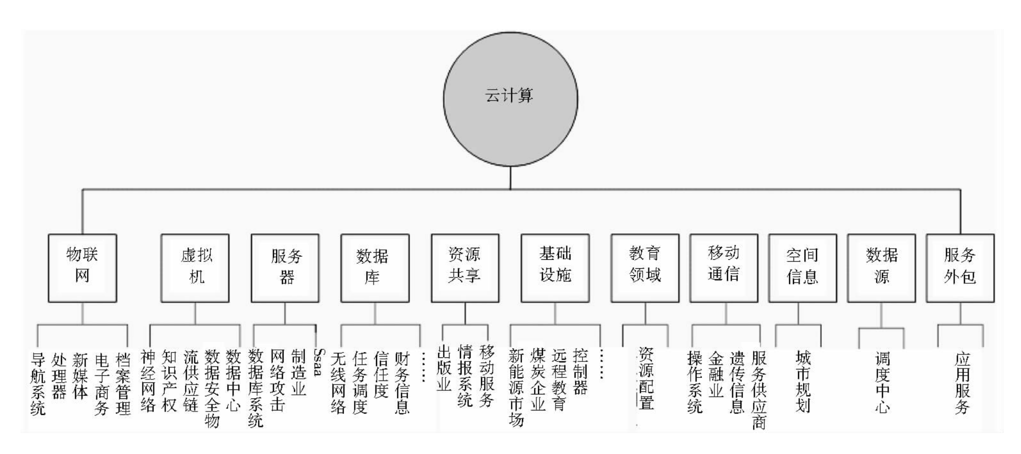


图 8　HLDA 方法－中文层次化知识结构

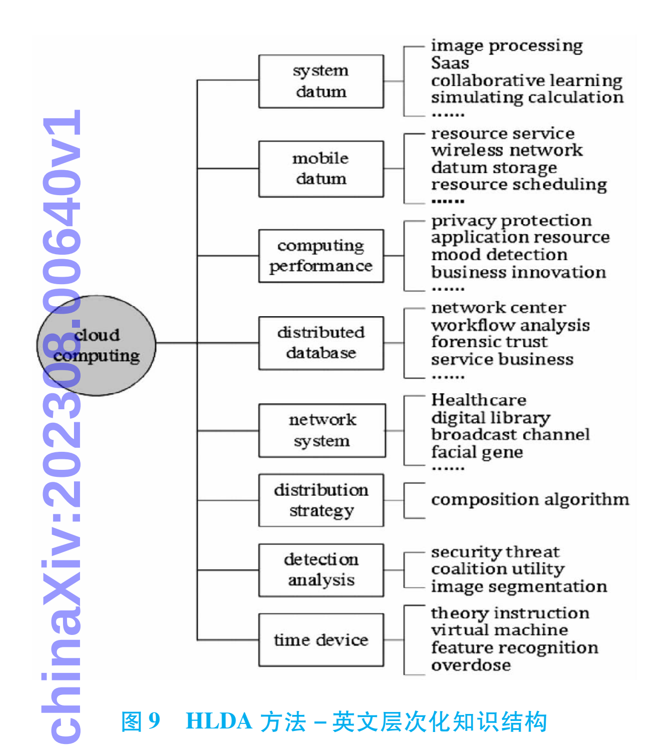


图 9　HLDA 方法－英文层次化知识结构

域的方方面面，全面且细致的将云计算领域分为十大研究方向，而 HLDA 方法得到的知识主题比较分散，且有重复交叉的地方，比如：虚拟机和服务器、数据库和数据源，且相对重要的数据安全、存储处理和算法方面均没有在知识结构中体现出来。以英文第二层知识结构为例，HSKSD 方法同样得到了结构关系良好，主题区分度明显的十大知识主题，相比之下，HLDA 方法得到的知识主题则不尽人意。

（2）层间主题继承性。层次知识结构发现与传统知识结构发现的主要区别就在于它是层次化的，通过对不同层的主题进行对比，观察各层间是否有较好的继承性，层次是否清晰。以中文为例，HSKSD 方法在第二层“算法优化”对应的第三层知识主题包括网络节点、资源分配、能耗分析、目标仿真、任务调度和负载均衡六个方面。HLDA 方法在第二层“物联网”对应的第三层知识主题包括档案管理、电子商务、多媒体、处理器、导航系统五个方面。对比发现，“算法优化”下的六个主题有较强的相关性和继承性，是隶属于“算法优化”的更细粒度的知识主题，而 HLDA 方法得到的“物联网”的下层主题继承关系并不明显，知识点较为分散。以英文为例，HLDA 方法得到的层次主题同样存在继承关系不强、甚至混乱的问题。

4.3　评价指标

本文的知识结构发现方法是一种无监督学习的方法，“云计算”领域的某些类标签或其它基准不适合作为参考，也不宜使用准确率、召回率、精度等传统的评价指标验证本文方法的有效性。因此，受到相关文献的启发，结合实验过程中的有益发现，本文综合采用以下 4 个评价指标评估 HSKSD 与 HLDA 方法的优劣：文档利用率 U（Utilization）、文档隶属度 M（Membership）、主题间独立性 I（Independence）[27] 和时间复杂度 TC（Time Complexity）[28]，并对前文未提及的符号做如下说明：① D_i 代表第 i 层的文档数量，i 代表层数，$i \in N+$；② D_T_{ik} 代表第 i 层主题 k 下的文档数量；③ $word_count$ 为预处理后词表中词的个数，即不重复词的个数。

各指标计算公式如下：

$$U_i = \left(\frac{\sum_{k=1}^{j} D_T_{ik}}{w}\right) \times 100\% \qquad \text{公式(1)}$$

$$M_i = Avg\left(\sum_{m=1}^{w}\sum_{k=1}^{j} P_{(D_{im},T_{ik})}\right) = \frac{\sum_{m=1}^{w}\sum_{k=1}^{j} P_{(D_{im},T_{ik})}}{w \times j} \qquad \text{公式(2)}$$

$$I_i = 1 - Avg(S_i) = 1 - \frac{2\sum_{k=1}^{j} S_{ik}}{j(j-1)} \qquad \text{公式(3)}$$

$$TC = \sum_{i=1}^{\infty} O_i(\text{niters} \times T_i \times word_count_i) \qquad \text{公式(4)}$$

4.3.1 文档利用率 在实验中,是否所有的文档都被分配到层次结构中是首先要衡量的一个指标,文档利用率 $U \geqslant 100\%$ 是合理的范围,在这种情况下,形成的层次结构是全面的、完整的,生成的主题才能覆盖所有的文档。

对于中文,HSKSD 方法的第二层由全部的 6 115 篇文献生成 10 个不同的主题,将每个主题下的文献相加是 10 374 篇,大于总数 6 115 篇,这是因为一篇文档可以属于不同的主题,只要其概率大于截取阈值 λ,就会被分配到相应的主题下组成文献子集,英文同理。HLDA 方法中,每一篇文档按路径寻找主题,路径唯一,其最终属于的主题也唯一,其实质是将所有文档按路径分配到不同的主题下,所以文献总和与文献集数量一致。

第二层各主题下的文档分布情况如下表 8、9 所示,根据公式(1)(i 取 2)可分别得到不同语料下的文档利用率 U。回顾表 7 比较参数设定中的阈值 λ,可以看出阈值大小与文档利用率成反比,即阈值越大,文档利用率越接近 100%,可以理解为,较高的阈值会让主题下被选中的文章数变少,从而降低文档被重复分配的比例。

表 8 中文第二层主题下文档数

文档数 D_T_{2k}	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	总计	文档利用率 U_2
HSKSD	1 188	627	880	675	1 296	1 037	1 211	758	1 391	1 311	–	10 374	170%
HLDA	560	669	399	1 252	224	1 955	107	712	108	85	44	6 115	100%

表 9 英文第二层主题下文档数

文档数 D_T_{2k}	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	总计	文档利用率 U_2
HSKSD	913	1 157	942	615	1 512	819	952	1076	1 073	524	9 583	198%
HLDA	244	2 800	540	623	292	28	169	147	–	–	4 843	100%

4.3.2 文档隶属度 一个主题下的所有文档及文档属于该主题的概率值 $P_{(D_{im},T_{ik})}$ 构成"文档－主题"矩阵,根据公式(2)(i 取 2)计算第二层主题下文献子集的平均隶属度 M,若 M 值较高,说明该主题下的文章具有较高的隶属度,聚类效果好。

从图 10 中可以看出,不同语料下 HSKSD 方法的 M 值均高于 HLDA 方法,说明我们的方法形成的主题对文档有更好的聚类效果。

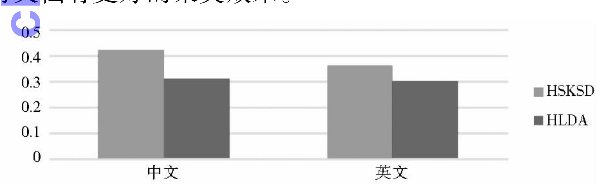


图 10 文档隶属度比较

4.3.3 主题间独立性 比较主题间独立性,实质是比较主题下的特征词之间是否存在差异。本文利用每个主题下的特征词根据公式(3)(i 取 3)计算第三层知识结构的主题间独立性 I,如果 I 值较高,说明主题间差异大,耦合度低,好的聚类效果也会提升主题表达性能和模型推广能力。

计算过程中,我们发现同一个主题的下层主题间相似性高,不同主题的下层主题间相似性低,符合一般聚类结果的类内高相似性和类间低相似性的目标,与实际情况一致。图 11 展示了不同语料下两种方法的主题间独立性。

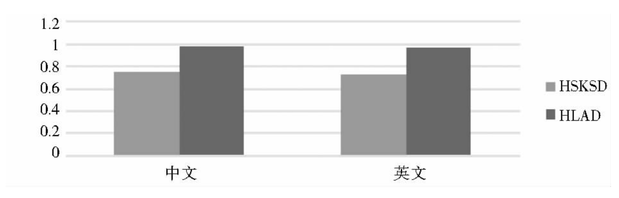


图 11 主题间独立性比较

从上图可以看出,HLDA 方法生成的主题间独立性较强,分析原因是 HLDA 在选择特征词的时候倾向于选择特殊的,出现次数少的,更具差异性的词语,而 HSKSD 方法则是选择文档中出现频率较高的、具有广泛代表性的词作为特征词,因此有不少特征词是一致的,这可能是造成主题间独立性较低的原因。

4.3.4 时间复杂度 算法的时间复杂度是反映算法优劣的重要指标,时间复杂度 TC 常用符号 O 表示,算法的时间复杂度越低,其效率越高。在实验环境为 MyEclipse 2015,JDK1.8.0,6GB 内存,Windows 7 操作系统的配置下,我们对算法的时间复杂度进行评估,见图 12。

每一层的复杂度为:$TC_i = O_i(\mathrm{niters} \times T_i \times D_i \times \bar{l}_i)$,其中 $\bar{l}_i$ 为各层文档平均长度,即 $\bar{l}_i = \frac{\mathrm{word_}count_i}{D_i}$,对该公式进行化简,即得到公式(4)(i 取 1,2,3),图 12 为根据公式(4)得到的时间复杂度计算结果,其中迭代

次数 niters 均为 1 000。

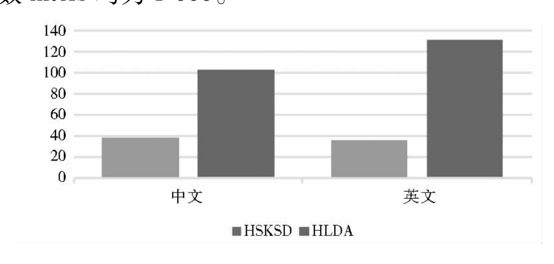


图 12　不同语料下的时间复杂度比较

观察上图，可以看到中文语料下，HLDA 方法的时间复杂度大约是 HSKSD 的 3 倍，而英文语料下则接近 4 倍，且 HLDA 方法有随层数加深，时间复杂度迅速递增的现象，所以在时间复杂度方面，HSKSD 方法明显优于 HLDA 方法。

4.3.5　综合比较　对前述比较标准进行综合评判，本文提出的 HSKSD 方法在文档利用率方面与 HLDA 方法均满足标准，文档隶属度和时间复杂度上则优于 HLDA 方法，HLDA 方法仅在主题间独立性上优于 HSKSD 方法，以上是从定量角度进行的比较。

此外，回顾 4.2.2 节中对知识结构直观的定性分析，单层次上，HSKSD 方法较 HLDA 方法发现的知识主题更加全面，少有交叉，主题间区分度较高。多层次间，HSKSD 方法较 HLDA 方法层次间继承性强，下层知识是上层知识的细粒度刻画，也可以解释为关联度或隶属性好。

通过以上分析，对本文提出的 HSKSD 方法和 HLDA 方法在 4 个定量指标和 2 个定性指标下进行综合评判，结果如表 10 所示。

表 10　不同方法的综合性能比较

综合评判	文档利用率	文档归属度	主题间独立性	时间复杂度	单层主题区分度	层间主题继承性
HSKSD	√	√		√	√	√
HLDA	√		√			

5　结论

本文提出了一种层次化的科学知识结构发现方法。在处理好的中英文语料下，首先利用 LDA 模型抽取“云计算”领域首层知识主题，然后设计了基于主题间平均相似性确定知识结构层次的算法和在“文档－主题”概率矩阵中自动筛选阈值确定下层文献子集范围的算法，结合这两大算法，可以帮助我们得到合理的知识结构层数和高质量的知识主题，然后对所有潜在知识主题进行映射，绘制最终的“云计算”领域知识结构层次树。通过对“云计算”领域层次树图的分析和解读，结合定量、定性指标，证明了我们的方法在文档归属度、时间复杂度、单层主题区分度和层间主题继承性方面较 HLDA 方法均有较大提升，也验证了本文方法的科学性和有效性。

抽取科学领域的知识主题并对其进行有效组织，对科研工作者和广大学者快速理解并掌握领域知识结构具有重要意义，现有知识结构发现方法涉及知识层次关系的不多，已有的层次化方法也存在知识结构难以控制，主题质量不佳等问题，本文提出的层次化科学知识结构发现方法，不但深入主题语义内部，给出了知识结构层次化的完整表示，还大大优化了知识结构发现过程，提高了知识主题质量，其知识结构构建速度更快，知识主题表征效果更好，这些优化具有一定的通用性，对其他领域科学知识结构的发现有一定借鉴价值。当然，本文的方法也存在一些不足，如较短的特征词使主题的可解释性不强，人工总结的主题代表词带有一定的主观性，衡量主题间相似性的方法有待提升等，未来笔者将在这些问题上进行探索和改进。

参考文献：

[1] 杜柏兰. 论档案人员知识结构的构建原则[J]. 北京档案,1998(7):28－29.

[2] 张振东. 情报人员“木”型学科知识结构[J]. 情报学刊,1991(6):475－478.

[3] CHARVET F F, COOPER M C, GARDNER J T. The intellectual structure of supply chain management: a bibliometric approach [J]. Journal of business logistics, 2008,29(1):47－73.

[4] 仲秋雁,曲刚. 知识管理学科知识流派划分及发展趋势研究[J]. 情报科学,2011(1):11－18.

[5] 宋歌. SNA 与 MSA 在揭示知识结构中的比较研究[J]. 图书情报工作, 2009, 53(8):106－109.

[6] 彭希羡,朱庆华,沈超. 基于社会网络分析的社会计算领域的作者合作分析[J]. 情报杂志,2013(3):93－100.

[7] ZHOU Z. Social network analysis of high cited authors based on domestic mapping knowledge domains[J]. Journal of modern information, 2012,32(8):97－100.

[8] RORISSA A, YUAN X. Visualizing and mapping the intellectual structure of information retrieval[J]. Information processing & management, 2012, 48(1):120－135.

[9] 李琬,孙斌栋. 西方经济地理学的知识结构与研究热点——基于 CiteSpace 的图谱量化研究[J]. 经济地理. 2014,34(4):7－12.

[10] RAVIKUMAR S, AGRAHARI A, SINGH S N. Mapping the intellectual structure of scientometrics: a co-word analysis of the journal scientometrics (2005－2010) [J]. Scientometrics, 2015, 102(1):929－955.

[11] KHASSEH A A, SOHEILI F, MOGHADDAM H S, et al. Intellec-

tual structure of knowledge in iMetrics[J]. Information processing & management an international journal, 2017, 53(3):705 – 720.
[12] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of machine learning research, 2003(3):993 – 1022.
[13] 王曰芬, 傅柱, 陈必坤. 采用LDA主题模型的国内知识流研究结构探讨:以学科分类主题抽取为视角[J]. 现代图书情报技术, 2016, 32(4):8 – 19.
[14] CHANG H C. The synergy of scientometric analysis and knowledge mapping with topic models: modelling the development trajectories of information security and cyber-security research[J]. Journal of information & knowledge management, 2016, 15(4):77 – 84.
[15] BLEI D M,GRIFFITHS T L,JORDAN M I,et al. Hierarchical topic models and the nested Chinese restaurant process[J]. Advances in neural information processing systems,2004,57(2):18 – 22.
[16] 陈静,徐波,王甜甜,等. 基于hLDA的图书内部主题层次组织研究[J]. 图书情报工作,2016,60(18):140 – 148.
[17] 陈亮,张静,张海超,等. 层次主题模型在技术演化分析上的应用研究[J]. 图书情报工作,2017,61(5):103 – 108.
[18] 张怡,邵裕东,张加万. 多源媒体文本主题演变的可视分析[J]. 计算机辅助设计与图形学学报,2017,29(12):2265 – 2272.
[19] 上海林原信息科技有限公司. HanLP[EB/OL]. [2017 – 08 – 10]. http://www.hanlp.linrunsoft.com/.
[20] The stanford natural language processing group. Stanford CoreNLP [EB/OL]. [2017 – 07 – 10]. http://www.nlp.stanford.edu/software/corenlp.shtml.
[21] 王鹏, 高铖, 陈晓美. 基于LDA模型的文本聚类研究[J]. 情报科学, 2015(1):63 – 68.
[22] 衡伟, 于佳, 李蕾,等. 应用hLDA进行多文档主题建模关键因素研究[J]. 中文信息学报, 2013, 27(6):117 – 128.
[23] BLEI D M, GRIFFITHS T L, JORDAN M I. The nested Chinese restaurant process and Bayesian inference of topic hierarchies [C]// International conference on neural information processing systems. Cambridge: MIT Press, 2012:17 – 24.
[24] ERL T, MAHMOOD Z, PUTTINI R,等. 云计算:概念、技术与架构[M]. 北京:机械工业出版社,2014.
[25] 朱敏. 云计算国内外研究现状综述[J]. 电脑知识与技术, 2015, 11(17):52 – 53.
[26] 王建冬, 刘洋, 王继民. 国内云计算研究领域核心作者群知识结构及演化路径分析[J]. 北京大学学报(自然科学版), 2013, 49(5):773 – 782.
[27] 李慧宗,周姣,王向前,等. 融合社会关系的用户标签主题模型[J]. 情报杂志,2017,36(3):165 – 172.
[28] 陈敏. 多模态语义知识库构造方法研究[D]. 武汉:华中科技大学,2014.

作者贡献说明:
李慧:设计研究方案,提出研究思路和论文修改意见;
田亚丹:负责实验实施,论文撰写和修改。

A Hierarchical Discovery Method of Scientific Knowledge Structure

Li Hui Tian Yadan
School of Economic & Management, Xidian University, Xi'an 710126

Abstract: [**Purpose/significance**] This paper proposes a new hierarchical discovery method of scientific knowledge structure, which provides reference for optimizing knowledge structure discovery process and improving knowledge organization form. [**Method/process**] Firstly, this paper constructed a hierarchical discovery method of scientific knowledge structure by using LDA topic model. Then, according to the average similarity degree among topics, it automatically determined the hierarchy of knowledge structure, and the literature subsets were intersected by filtering threshold automatically in the "document-topic" probability matrix. Finally, it adopted tree diagram to display the science knowledge structure and explore the correlation and inheritance of knowledge points. Besides, we also compared our method with HLDA method which is a hierarchical topic model. [**Result/conclusion**] The result shows that the knowledge structure obtained by our method is better, the representation of knowledge topic is stronger and it has the higher operation efficiency. In addition, compared with the HLDA method, our method has a great improvement on the topic differences of the single layer and the topic inheritance between layers.

Keywords: LDA cloud computing hierarchical knowledge structure